SESUG 2020 Paper 157

Application and comparison with several types of Ensemble Algorithms

Yida Bao, Dr. Philippe Gaillard, Auburn University

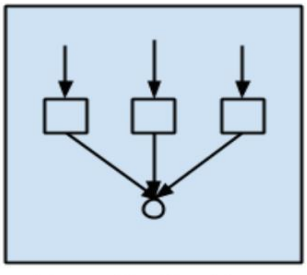
ABSTRACT

Ensemble method is one of the common algorithms in Machine learning field. Researcher use algorithms to generate models composed of multiple weaker models that are independently trained and whose predictions are combined in some way to make the overall better prediction. In this paper, we will focus on Gradient Boosting Machines(GBM), Stacked Generalization (Stacking) and Random Forest those three typical algorithms. We use different algorithms to test different types of data set and compare the accurate result. **SAS®** Enter Miner would be used to achieve this process.

INTRODUCTION

Machine learning is a data analysis technology that allows computers to perform the innate activities of humans and animals: learning from experience. Machine learning algorithms use computational methods to "learn" information directly from data, rather than relying on predetermined equation models. When the number of samples available for learning process increases, these algorithms could adaptively improve performance. Recently, Machine learning is one of the fanciest statistical methods; it has been widely used in a lot of area, such as data mining, computer vision, natural language processing, biometric identification, search engines, medical diagnosis, stock market analysis, speech and strategic games, and robotics.

In this paper, we will focus on dealing with several categorical variables. Classification method would be applied in the research. Classification stands an extremely vital position among statistical methods. The idea of trying to design and build a machine that recognizes difference modes would be the pursuit of science researchers. Many applications, from automatic speech recognition to fingerprint recognition, optical character recognition, DNA sequence analysis, etc., clearly demonstrate the essential role of a reliable and accurate pattern recognition system. When facing complex sensor data, classification is the first key processing step when extracting useful information for all intelligence systems.



Ensemble Algorithms

Ensemble algorithm will be used to build the model. Ensemble methods are models composed of multiple weaker models that are independently trained and whose predictions are combined in some way to make the overall prediction. Much effort is put into what types of weak learners to combine and the ways in which to combine them. This is a very powerful class of techniques and as

such is very popular.

ENSEMBLE METHODS

Ensemble learning would improve misclassification rate results by aggregating multiple weighted models. Dietterich explained three fundamental reasons for the success of ensemble methods: statistical, computational and representational. In addition, bias-variance decomposition and strength-correlation also explain why ensemble methods work. Ensemble learning are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance, bias, or improve predictions. This approach allows the production of better predictive performance, especially compared to a single model. That is the reason that ensemble methods placed first in many prestigious machine learning competitions, such as the Netflix Competition, KDD cup, and Kaggle.

Ambiguity decomposition only applies to a single dataset with ensemble methods. For multiple datasets, bias-variance covariance decomposition is introduced [12]–[15] and the equation is shown:

$$E[f-t]^2 = \text{bias}^2 + \frac{1}{M}\text{var} + \left(1-\frac{1}{M}\right)\text{covar}$$

$$\text{bias} = \frac{1}{M}\sum_i (E[f_i] - t)$$

$$\text{var} = \frac{1}{M}\sum_i (E[f_i - E[f_i]]^2$$

$$\text{covar} = \frac{1}{M(M-1)}\sum_i\sum_{i\neq j} E[f_i - E[f_i]](f_j - E[f_j])$$

where t is the target and f_i is the output from every model , M is the size of ensemble. From the equation, we can see the term *covar* can be negative, which may decrease the expected loss of the ensemble while leaving bias and *var* unchanged. Beside the *covar*, the number of models also plays an important role. As it increases, the proportion of the variance in the overall loss vanishes whereas the importance of the covariance increases. Overall, this decomposition shows that if we are able to design low correlated individual learners, we can expect an increase in performance.

name	DEFINITION
Adaboost	ADAPTIVE BOOSTING
WEIGHTED AVERAGE	WEIGHTED AVERAGING OF THE POSTERIOR PROBABILITIES
GBM	GRADIENT BOOSTING MACHINES
GBRT	GRADIENT BOOSTED REGRESSION TREES
HP FOREST	RANDOM FOREST

Table 1 : Nomenclature

DATA DESCRIPTION

In this paper, we will use three different dataset to verify the result. Each dataset variable has its own target variable. We will use different model which act as classifiers to reach a high misclassification rate. Below we will carry out a brief data description.

- **Sports Article** contain 1000 sports articles in the dataset, including 635 articles label as objective labels and 365 articles label as subjective. Based on different types of features, we got 48 variables.(dependent variable)

- **Diabetes** contain 769 people's blood inspection data, including 500 people with diabet negative label and 365 patient with diabetes positive label. We got 8 variables.(dependent variable)

- **Breast Cancer** contain 117 people data in the dataset, including 53 people label as negative labels and 65 patient label as positive. We got 9 variables.(dependent variable)

MEHODOLOGY

We will use SAS Enterprise Miner platform instead of SAS 9.4 to deal with the sports article dataset. SAS Enterprise Miner helps us analyze complex data, discover patterns and build models so we can more easily detect fraud, anticipate resource demands and minimize customer attrition.SAS Enterprise Miner offers many features and functionalities for the business analysis to model the data.

We build a SAS Enterprise Miner diagram for the ensemble model.

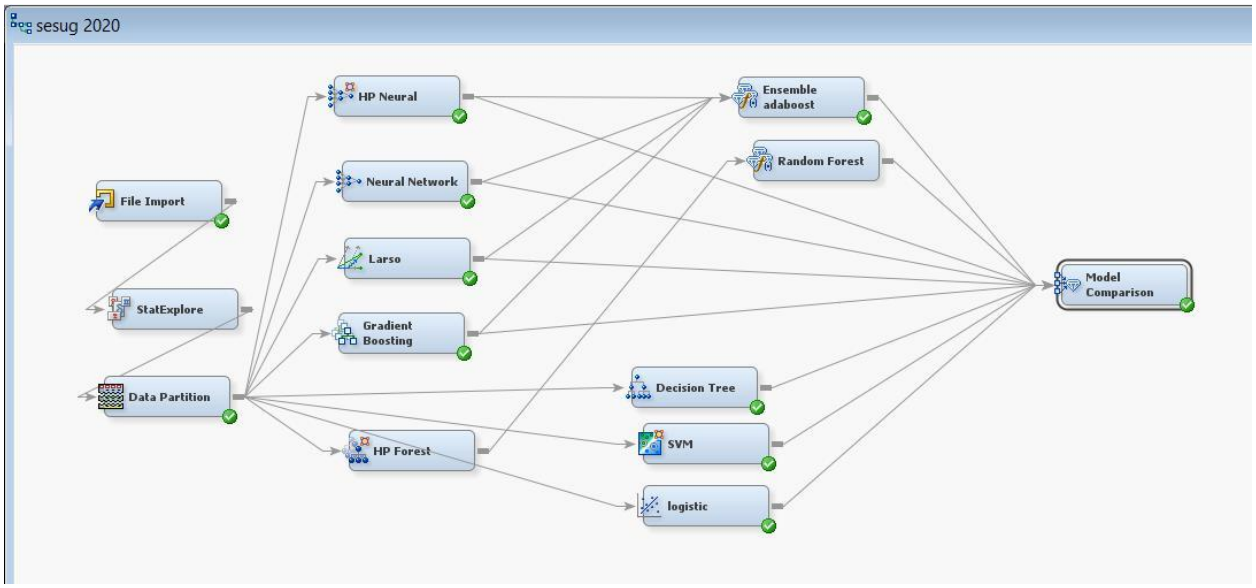


Figure 1. SAS Enterprise Miner Diagram

- Our first step is to clean and split data in SAS software. It's simple to use **File Import Node**, and then use **Data Partition Node** to split the data into three part, Train, Test.

  It is some data sets that we already know the input and output to train the machine to learn, and find the initial parameters of the model through fitting. We ensure a considerable proportion of train parts. For validation part, it is also some data sets for which we already know the input and output. By allowing machine learning to optimize and adjust the parameters of the model.

  Training Set In traditional machine learning, the general ratio of these three is training, validation, test is equal to 50/25/25. In our paper, the ensemble model does not require a lot of adjustments, we use the ratio **training/test = 7/3**.

- We have introduced almost all common machine learning methods, including neural Network, lasso regression, GBM (GRADIENT BOOSTING MACHINE), Random forest, for these methods, we separate these methods into two ensemble models :

  sequential :ensemble methods where the base learners are generated sequentially. The basic motivation of sequential methods is to exploit the dependence between the base learners. The overall performance can be boosted by weighing previously mislabeled examples with higher weight.

  parallel : ensemble methods where the base learners are generated in parallel (e.g. Random Forest). The basic motivation of parallel methods is to exploit independence between the base learners since the error can be reduced dramatically by averaging.

- We will use Model Comparison Node to compare different model, and find out the performance for different dataset.

RESULT

	Breast Cancer	Diabetes	Sports Article
Lasso Regression	41%	21.98%	17.33%
SVM	36%	21.55%	15.33%
Logistic Regression	35%	22.4%	19.33%
GBM	22%	22.8%	14.7%
Neural Network	31%	20.8%	18%
Sequential Ensemble	27%	20.6%	14.66%
Parallel Ensemble	36%	25%	18%

Table.2 Misclassification Rate for different datasets

We use SAS enter miner to quickly get the results we need. Among the final results, we are most concerned about the misclassification rate of test part. The lower the misclassification rate, the better represent the performance of the classifier.

From the results, the results of the ensemble algorithm seem to be very decent among so many machine learning algorithms. This also seems to be reasonable, since the ensemble algorithm combines several models. Before going into a more in-depth discussion, we need a more precise comparison.

	Breast Caner		**Diabetes**		**Sports Article**	
	SEQ Ensemble	PAR Ensemble	SEQ Ensemble	PAR Ensemble	SEQ Ensemble	PAR Ensemble
Lasso	+51.8%	+13.89%	+6%	-12.08%	+18.2%	-3.7%
SVM	+33%	+0%	+4.6%	-13.8%	+4.5%	-1.5%
Logistic	+3%	-2.78%	+8.7%	-10.4%	+31.86%	+7.3%
GBM	-18.5%	-38%	+10.6%	-8%	+0.2%	-18.3%
Neural	+14.8%	-13.88%	+0.97%	-16.8%	+22.8%	+0%

Table.3 Misclassification improvement

From the above classification improvement table, the advantages of Sequential Ensemble algorithm can be clearly found. In some parts, the amount of data for Breast cancer is so small that the ensemble algorithm is 51.8% better than Lasso regression, which seems impossible. In each data set, the Sequential Ensemble algorithm has shown better performance than most other algorithms

On the other hand, Parallel Ensemble algorithm does not perform as good as sequential ensemble, from the construction of the algorithm, we think this is more due to the adaptability of the different data set

DISCUSSION

Ensemble methods would help us to get a relative better misclassification rate results by aggregating multiple weighted models.

The research in this paper showed that several models and their ensembles tend to have similar fit statistics result and showed that it is useful to evaluate the performance by checking the misclassification rate .

According to our results, just as other model, some single model will even perform better than ensemble algorithm. Different data form, different train /validation/test ratios will affect the final result. Which classification method is the best? it seems that there will never be a certain conclusion. The researcher must consider various situations and conduct different forms of experiments to determine the model they use.

REFERENCES

- *Getting Started with SAS ® Enterprise Miner 7.1*
- *"An introduction to statistical learning", Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani.*
- *"Pattern Classification" , Richard O. Duda*
- *Ensemble Learning to Improve Machine Learning Results, Vadim Smolyakov*
- *Ensemble Modeling: Recent Advances and Applications, Wendy Czika, Miguel Malgonado, and Ye Liu, SAS Institute Inc.*
- *Ensemble Classification and Regression – Recent Developments, Applications and Future Directions, Ye Ren, Student Member, IEEE, Le Zhang, Student Member, IEEE, and P. N. Suganthan, Fellow, IEEE*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

YIDA BAO, Math PHD Candidate, SAS certified advanced programmer
Department of Mathematics and Statistics, Auburn University
E-mail: yzb0010@auburn.edu

Philippe Gaillard, Associate Professor
Department of Mathematics and Statistics, Auburn University
Director of Statistical Consulting Center
E-mail: prg0007@auburn.edu